DOCUMENT RESUME

ED 380 406 SP 035 640

AUTHOR Talbot, Gilles L.
TITLE Revitalizing Teacher-Made Tests: Quality Control Procedures.
PUB DATE Sep 94
NOTE 28p.
AVAILABLE FROM G. L. Talbot, 790 Neree Tremblay St., Ste-Foy, Quebec, Canada G1V 4K2 ($10 Canadian).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Students; Foreign Countries; Higher Education; Quality Control; Teacher Developed Materials; Teacher Improvement; *Teacher Made Tests; *Test Construction; Test Format; Testing; Test Items; Test Reliability; Test Results; Test Theory; *Test Wiseness
IDENTIFIERS *Colleges of General and Professional Education PQ; *Quebec

ABSTRACT

This paper offers college teachers guidelines for improving their teacher made tests. It notes that teachers may focus on how well students have learned course objectives while being unaware of how the testing process itself contributes to the results obtained. The paper reports the results of a test-taking workshop designed to improve college students' testing awareness and test taking skills. An opening section identifies eight steps in the test construction process and discusses item bias and analysis, fairness in grading, and motivations for testing. The second portion of the paper describes a workshop to teach college students test taking skills through a variety of sample activities. Analysis of these activities allow demonstration of the following testing indexes: sensitivity for guessing index, instructional index, discrimination index, and difficulty index. Other testing concepts introduced by examples from the workshop activities include instrument bias, test validity, and interpretation bias. The institutional and social context of test administration and construction is also addressed. A conclusion notes that workshop participants improved their attitudes about tests and appeared to realize the relationship between real effort and improved results. (Contains 29 references.) (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Revitalizing Teacher-Made Tests: Quality Control Procedures

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E Talbot

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

ED 380 406

by

Gilles L. Talbot
Champlain-Saint Lawrence
Ste-Foy, Province of Québec
Canada G1V 4K2

September, 1994

SP035640

BEST COPY AVAILABLE

POINTS OF VIEW OR OPINIONS EXPRESSED IN THIS REPORT ARE THOSE OF THE AUTHOR AND THEREFORE DO NOT NECESSARILY REPRESENT OFFICIAL GOVERNMENT OR COLLEGE POSITION OR POLICY.

Copies of this report may be ordered from the author, at the following address, and for a nominal fee of ten dollars (10$) for photocopying, mailing etc.

Gilles L. Talbot
Champlain-Saint Lawrence
790, Nérée Tremblay Street
Ste-Foy, Province of Québec
Canada G1V 4K2

Revitalizing Teacher-Made Tests: Quality Control Procedures[1]

Scarvia Anderson (1987), in a constructively reflective examination of the relationship between teacher-made tests and higher education, has observed:

> Teacher-made tests are more than assessment devices: They are a fundamental part of the educational process. They can define instructional purposes, influence what students study, and help instructors to gain perspective on their courses. How well the tests accomplish these purposes is a function of their quality.

We tend to agree with the following generalizations, or policy statements, made by this former senior administrator for Educational Testing Service:

> What college teachers should be called on to defend is the quality of the tests that they give and the influence that the tests exert on student learning.

There is an urgency to this appeal. Teacher-made tests, whether personally developed or taken from test banks that accompany textbooks, would appear to be notoriously incomplete, perhaps inadequate. Even those who set the standards for testing in education fail to put them into practice. Ellsworth, Dunnell and Duell (1990) examined the difference between what authors of educational psychology textbooks were proposing in their books as guidelines for building sound multiple-choice tests with actual test banks accompanying their texts. "The results indicated that approximately 60% of the items violated one or more guidelines" (page 289). If the people who write to set the standards do not even bother to check their own tests what is to be said for the training and motivation to ensure quality control? The situation that prevails in the Province of Québec's college system is probably reflective of the problem:

> Par ailleurs, évaluer les apprentissages reste, le plus souvent, un geste privé. Chacun construit ses instruments d'évaluation et effectue ses corrections dans la plus grande discrétion. Bien audacieuse la direction des études qui demanderait d'examiner, en même temps que les plans de cours, les instruments d'évaluation des apprentissages...(Association québécoise de pédagogie collégiale, page 7)
>
> (Besides, the evaluation of learning remains, most often,

---

[1]A cordial thank you to my colleague Chris Vandenberg, Professor of English and English Literature at our college, who proofread this article. I find reassurance and motivation in his comments, devotion to his students and concern for quality in higher education.

> a task carried out in private. Each person designs tests and makes corrections under very private conditions. It would be a very daring Dean who would ask to review the course outlines in conjunction with proposed testing to evaluate student learning (Association québécoise de pédagogie collégiale,1994; page 7)

In fact a strong and persistent illusory correlation seems to exist between what and how teachers think they test, with what and how they do test. The illusion arises as part of the process in which teachers attempt to learn how well students have learned course objectives while ignoring how the testing itself may also contribute to the results. Student test scores are supposed to reflect knowledge acquired from instruction and other pedagogical activities. The symptoms of quality control procedures left unattended are instructional, instrument, interpretation and item biases. These biases encourage students to develop strategies based on testwiseness, guessing and cramming. Teachers are responsible not only for what students learn but also how and when they learn it. Pintrich et al.(1986) describe this responsibility in a field called instructional psychology which "...aims for the development of motivation, cognitive structures, and a repertoire of skills and strategies for learning and problem solving" (page 612).

This paper proposes to help teachers not only with the content knowledge (what to do) but also to afford them the same concern, as we expect them to extend to students, for contextual and procedural knowledge about tests (i.e., when and how to revise tests). A special concern is the effects these procedures have on motivation.

Revitalizing Teacher-Made Tests

Testing the test

Revitalizing teacher-made tests rests on several assumptions. First are the general background considerations: we assume tests are planned and not prepared at the last minute; that teachers grade fairly, and not politically, as in protecting one's reputation by artificially inflating or deflating grades.

Second, test-planning and test-taking performances are related to the frequency, spacing and weighing of tests which are in turn related to formative feedback *possibilities*. The frequency, spacing and weighing of tests make statements about how the teacher views formative feedback. Will there be opportunities for students to receive formative (i.e. diagnostic) feedback and time to make necessary changes? Will students have realistic opportunities to demonstrate they have corrected deficient behaviors?

Third, we assume there exists time, opportunity, and a willingness for students and teachers to invest in the student-teacher relationship. In other words, the contingency between real effort and results in working together means that each will do as much as s/he can, and not necessarily as much as the student expects of the teacher, or vice versa. In this context teachers need to examine their own expectations for student performances. Not every student will share the teacher's enthusiasm and need for understanding. So, the "control" aspect, in "quality control", means that each party strives to do his/her best, within the limits of one's abilities and motivations for achievement. If teachers are to expect students to rise to their potential, then students also may expect teachers to do the best they can.

Stiggins, Conklin and Bridgeford (1986) have recognized the importance of "quality control" of teacher-made tests in these terms:

> Teachers should be provided with relevant, focused preservice and inservice training in classroom assessment strategies and useful quality control procedures. At least some of the content of that training is suggested by the research reviewed here (page 15).

Even though many college teachers may not have taken pre-service courses or received training, the suggestion is made that it is possible to grow out of this limitation by engaging in professional development for "inservice training in classroom assessment." Apparently research driven conclusions showing the pressing need for such changes in college faculty were not obvious enough. Stiggins (1988) re-emphasized it in "Revitalizing Classroom Assessment: The Highest Instructional Priority":

> ...we are again failing to address the central issue in school assessment: insuring the quality and appropriate use of teacher-directed assessments of student achievement used every day in classrooms from coast to coast. Classroom assessments are the tools teachers use to determine whether the pace of instruction is appropriate and whether their instructional objectives have been met. They are the assessments that determine to large extent what grades students receive and how they perceive their school experiences (pages 363-364).

Stiggins recognizes the motivational importance that grades and grading, and thus tests, have on students. More importantly he is working to increase teacher awareness of this reality by appealing to them to engage in analyses of their tests. What has not been so clear in the literature is how to use the results of item or

content analyses[2] to help teachers "gain in perspective" about their courses and tests, and also to help students move towards a learning orientation. We believe that actively involving students in testing has strong inputs into their motivational beliefs. As Cross (1988) has put it: "It is one thing to recommend that students be actively involved in learning; it is quite another to suggest to teachers how that can be accomplished (page 15)."

It is hypothesized that students and teachers interacting over the results of item or content analyses would produce gains to both students and teachers about how to move towards a learning orientation and fair testing/grading,respectively. The "results of item analyses" are operationally defined as the processes and results which identify any instrument, item, interpretation and/or instructional biases. The "learning orientation" focused on testing the test and, indirectly on examining how instructional objectives have been met.

How teachers may perform steps 5 and 8 are the basis of the quality control procedures for revitalizing teacher-made tests. Testing the test then, means that teachers must be aware of the following steps:

1) test planning;
2) test construction;
3) test development;
4) test reporting;
5) review of steps 1 through 4;
6) test-taking and test administration;
7) test scoring and grading; and,
8) test verification of the results
   8.1 comparing results obtained in step 7 with processes to build the test in step 5; and,
   8.2 assessing attainment of teaching/learning objectives in step 7) with test planning in step 1).

---

[2]"The technique of item analysis is then used to sort through the data in order to select the best statements for the final scale. It consists of calculating the extent to which the responses to individual statements are correlated with the total score. ... Statements that correlate with the total score are then chosen for the final scale. The rationale is that statements that have low correlations witn the total score will not be good at discriminating between ...respondents." (Rosenthal,R. and Rosnow,R.L. (1991) Essentials of Behavioral Research - Methods and Data Analysis,2nd ed., Montréal: McGraw-Hill, page 196.

"Content analysis involves the objective, systematic strategy of decomposing messages and then evaluating and classifying their contents in order to reveal their specific characteristics." (Rosenthal and Rosnow, 1991; page 171). In this context, content analysis is equivalent to first step of the scientific method: systematic and accurate observations of a phenomenon.

An instrument or item bias refers to the fact that the whole test ("instrument") or a particular question or answer ("item") was not related to the learning objectives, had poor discrimination and/or difficulty indexes, and many even have had contextual or grammatical problems in directions, vocabulary or either the question or answer stems. An interpretation bias refers to the fact that the question or answer stems may be interpreted in more than one way. The same holds true for instructional biases except that faulty interpretations may have arisen during the delivery of the lectures or a reading of the assigned materials. In such cases the teacher failed to see these possible alternative interpretations and to deal with them appropriately in the lecture. So, if the problem is with interpreting a question because of inherent ambiguities it is an interpretation bias. If the interpretation arose earlier with respect to learning objectives, then it is an instructional bias.

Thus, the instrument and item biases are inclined to be objective while the interpretation and instructional biases tend more towards subjectivity. By working with students on the development of the discrimination, difficulty, instructional bias,and guessing bias indexes one increases interactions and mutual self-awareness about how the teacher and students know if the test is doing its job. The process of quality control of tests takes on motivational properties through this formative feedback.

Cross (1988) has suggested that setting a smaller and more realistic goal gets both students and teachers on the path towards such a "learning" orientation:

> If teachers had better feedback from students about how students think about themselves as learners, teachers might modify their teaching. Understanding <u>how</u> students engage themselves in the process of monitoring their own learning, for example, should be helpful to teachers in making assignments and designing class experiences to encourage active student involvement. (page 15).

Careful and full exploitation of the potential of item and content analyses is an approach that avoids having to create still more tests and assignments to provide ungraded formative feedback. Teachers avoid repeating mistakes by re-assigning work or tests and to have to correct still more work which will not count towards the course grade. Also, this change is in keeping with changing institutional policy to de-emphasize grades instead of learning by providing more formative and less summative feedback and evaluations. Also, the approach provides objective inputs into fair grading practices.

Fairness in Grading

Pollio and Humphreys (1988) have succinctly summarized the motivations, fairness in grading issues, and quality control functions of teacher-made tests:

> We want to argue that we should do everything possible to ensure that grades and grading enrich the academic setting, facilitate fruitful interactions between instructors and students, and serve to augment rather than impede the course of college learning (pages 85-86).
>
> Most faculty have had little training in constructing such test items, and too few discuss classroom testing with colleagues, even though there is much evidence ... to indicate that what and how we test is a powerful force in shaping what and how students learn (page 87).

Students are painfully aware of the effects the grades and grading practices have on their motivation for learning (Pintrich and Johnson,1990). Students perceive teachers to be oblivious to the effects that teachers' grading attitudes and practices have on students` attitudes and motivations for tests. These problems are not unrelated one to the other and they would seem to be a logical topic of discussion in the context of student-teacher interactions. Teachers would learn and teach students how to relate feedback to learning orientation (i.e. contingency of effort to results). Students could be more motivated to move towards a learning orientation (i.e. study skills and learning strategies) if they could see that teachers are willing to reciprocate by considering how tests motivate student learning and grades.

Motivations

Cross (1988) has also addressed the issue of testing when she writes about "assessment-for-improvement":

> If the ultimate purpose of assessment is to improve teaching and learning, then the results of a successful assessment must eventually bear directly on the actions of teachers in their classrooms (page 1)
>
> ...the role of feedback in the assessment-for-improvement model is to provide a continuous flow of information that is useful in shaping the process of teaching and learning while it is in process. This is generally referred to as "formative evaluation," and is most effective if it is <u>not</u> made public and emphasizes competencies instead of comparisons (page 2)

We believe college teachers are willing to respond to practical suggestions for improvement. The crux of the problem, as Cross recognizes it, is that assessment-for-accountability is after the fact, goal-oriented, and favors comparisons amongst teachers. In this context it is understandable that teachers have perceived formative evaluations as threats and reacted accordingly. Formative evaluations are perceived primarily as the appraisals of our personal and social inadequacies. Formative evaluations strive to answer "why" type questions. They also suppose that teachers have the insights, familiarity with constructs[3] and relevant vocabulary, and the strength to perceive the situation as an opportunity to grow rather than to justify one's actions. For these reasons, we feel that formative evaluations are more "evaluative" than "formative."

Formative feedback, concerned with process and competencies, appeals to "when" and "how" type questions. The answers to such questions favor the continuous shaping process and focus on the task rather than the goal. In so doing, we focus on changes to be made to the task which are far less threatening than what we perceive to be changes to be made to the teacher. Teacher-made tests, in this context, should benefit more with formative feedback than with formative evaluations. What we have observed is that there are many excellent sources of information for the development of teacher-made tests which rarely get the attention and circulation they deserve (see, for example: Taylor,1978; Rubadeau et al. 1990; and, Williams, 1991). We suspect, from our own in-house work with peers, that these excellent documents are threatening. Teachers' usual reply are of the type: "Why should I?", or "What is wrong with the test?", instead of "How could the test...?", or "When should the test...?"

There is a very close parallel in the suggestions made by Cross in the way she uses "assessment" in referring to teacher evaluations, and to the application of these suggestions, as we maintain , to the idea of revitalizing "teacher-made tests":

> A related contrast that I should like to draw in setting the stage for a discussion about using feedback more effectively in assessment is a distinction between what I shall call "direct" and "indirect" models of assessment (page 2).
>
> Classroom teachers are directly involved in instruction. Through their <u>own</u> actions, they can change the quality of teaching and learning in the classroom. For that reason alone, it is important to get teachers <u>as individuals</u>

---

[3]Construct: "An abstract variable, constructed from ideas or images, which serves as an explanatory term." (Rosenthal,R. and Rosnow, R.L. (1991; page 616)

> involved in this assessment movement (page 2).
>
> Research on teaching and learning is moving in the direction of studying cognition and learning in the context of the subject or content taught, and we need the participation of discipline-oriented faculty in assessment and research on teaching and learning so that we may know how to improve the process (page 3).
>
> My hypothesis is that the most effective form of assessment is one that is continuous, that occurs as close as possible to the scene of the action in teaching and learning (the classroom), and that provides diagnostic feedback to both teachers and students -- to teachers on how they can improve their teaching, to students on how they can improve their learning (page 3).

Revitalizing teacher-made tests fits this mandate quite neatly. It is, in practice, the "highest instructional priority."

> Feedback to students on tests and other assignments should contain more than the simple notation of a grade or indications of what is right or wrong. A test is not only a grading device but also a teaching technique in its own right (Milton,1982;page 95).

Pintrich,Marx and Boyle (1993) have presented a scholarly argument to show the merits of our proposition: "Students follow as they are lead!" If teachers want students to engage in self-reflected learning then teachers must engage in self-reflected teaching. Preaching self-reflective awareness in others without participating is not conducive to motivating others to change.

> There is abundant anecdotal evidence that much of what happens in school is driven by need to maintain bureaucratic and institutional norms rather than scholarly norms. Much research literature documents this interpretation; it is likely that many students hold similar views of schools and the instructional activities that take place there. To the extent that this is true then, it is unlikely that individual conceptual change will take place without restructuring classrooms and schools along lines that will foster the development of a community of intentional, motivated, and thoughtful learners (page 193).

Theoretical Framework

Motivated cognitions "...refer to the complex interplay in which cognition is at once the servant of motives ... and also the planner and clarifier (Covington, 1983;page 140)" which are the bases for reflective self-awareness. It is quite likely that such "reflective practices" (Argyris and Schön,1974; Schön,1983) may be used to help us understand the thoughts, feelings, behaviors and motives behind both academic achievement motivation in students, and revitalization of classroom tests by teachers. Kirby and Teddlie (1989), who have derived the reflective teaching instrument from Argyris and Schön's work, define the Personal Causation scale in this way:

> The third requirement for developing an effective theory of practice is personal causation. The practitioner must be committed to the personal and professional values in setting the problem and must accept responsibility for actions taken. Unless there is a strong commitment to values and to self, the practioner will be unable to question the conventions of the profession when necessary, will have difficulty admitting perceived failure, and will resist testing.
>
> Personal causation is the acceptance of responsibility for actions and their consequences. (page 46)

The inability "...to question the conventions of the profession when necessary..." is strongly maintained in place by the availability heuristic[4] and both the cognitive and behavioral confirmatory biases. The *availability heuristic* is derived from our impressions that information is objective and accurate because it is readily and repeatedly made available to us. Teachers assume tests are doing their job: until too many students complain or fail; because the teacher's practices have been sanctioned by the administration as a result of formal evaluations; because the test banks accompanying textbooks propose questions that meet teachers' course objectives; and, because the multitude of questions in the test bank allow teachers to choose definitional, application or interpretive type questions for each learning objective of the chapter. All of these evaluations feed into the availability heuristic. Few teachers attempt to see beyond the evaluations to ascertain instrument, instructional, item or interpretation biases.

---

[4] Heuristic: "A rule of thumb that guides problem solving but does not guarantee an optimal solution. Heuristics are often used as shortcuts in solving complex problems." Wade,C. and Tavris,C. (1993) Psychology 3rd ed. N.Y.: HarperCollins page G-5.

When we tend to notice and process information to support our choices (*cognitive confirmatory bias*) or we engage in behaviors to meet the expectations we think others hold of us (*behavioral confirmatory bias*) we are working, unaware, with perceptual and memory processes that are both selective and constructive. Teachers, as with all groups of people, strive towards internal and external consistency. They want their public and private self-images to be similar and credible. It is no wonder then that teachers operate to process information that minimize internal/external dissonance and that provides normative influences to the self to conform to public self-images.

Is it any wonder then that to merely suggest quality control of tests to teachers leads them to the reply: "And what's *wrong* with my tests?" Evaluations have become so commonplace as part of the assessment-for-accountability process that we tend to see it everywhere. Teachers have to make a special effort to recognize, and to respond calmly, to the rather logical possibility that their tests could do with improvement through formative feedback and <u>not</u> through formative evaluation. "Change" has come to connote "evaluation" and "assessment-as-accountability" so much that we fail to see anything to do with change as a source of feedback for improvement.

The theory of reflective practices (especially reflective awareness and personal causation) would contribute to reflective teaching. Students and teachers could monitor <u>individual</u> efforts. As the levels of shame (differences between ideal and actual roles), guilt (not attaining the ideal role levels), and doubt (one has the ability to meet expectations for ideal roles) subside, the availability heuristic and the cognitive and behavioral confirmatory biases also relax. We suggest that as this happens the negative internal or external motivational states will give way to positive motivational experiences for teachers. Such states as pride, satisfaction, mastery and confidence (internal) or praise, attention and other forms of recognition (external) from one's students, peers and, hopefully, the administration.

We hypothesize that knowledge about instrument, instructional, interpretation and item biases provides formative feedback to teachers about teaching and to students about learning. Information about these biases contributes to reflective teaching and learning orientations. Revitalizing teacher-made tests affects student learning and the student-teacher relationship to the degree that students and teachers learn the answer to a very important question: "How well do teacher-made tests account for the quality of teaching and learning?" Perhaps then students will also become more receptive (i.e. less threatened) to teacher's concerns with: "How much real effort are you putting into this course?"

Subjects

The Québec college system operates on an open-door basis. After six years at the elementary level, five years at the secondary level, and just before the 3 years of discipline specific intensive studies in University, students pass through 2 to 3 years of college. Those who opt for technical or vocational programs terminate after 2 to 3 years of college study. The brightest of college students usually seek admission into the 4 year M.D. or Bachelor of Laws programs. All studies are paid for to the end of high school. The Province of Québec also underwrites most of college educational expenses and a good part of University studies. If I interpret the policy correctly, the intention is to help each person attain his/her potential without regard to race, religion, sexual orientation, political beliefs or financial limitations.

Many Cégep students are notoriously poor in vocabulary, reading rate and comprehension, taking and revising lecture notes etc. Many students also do not have appropriate study skills or learning strategies (Bateman, 1989). The interaction of the two in one person may account for the relatively high abandon and failure rates (about 35%). Of course, there are many other factors such as part-time employment, social life, extra-curricular activities and sports. Whatever the motives and causes for academic success and failure, it remains that our Cégeps (that's an acronym for the Québec College system) were conceived, and the faculty appointed, to give everyone an equal opportunity. To this end, our Cégeps are concerned with developmental education which means declarative knowledge (what to study) should also be accompanied with procedural (how to study) and conditional knowledge (when and where to study). The formidable task of academic advising is, in the context of a philosophy of developmental education, to get students to make real effort (Talbot,1994).

The literature on academic achievement motivation has shown that apparently many low ability students are actually quite diverse. Some, of course, do not have the intellectual skills. Some are caught up in personal,family or financial problems. Still others are too busy in the pursuit of adolescent gratifications to care about their future. However an appreciable minority (10%-15%) are too threatened by the possibility of failure ("ego-oriented"), or too extrinsically motivated (they don't see learning but only grades as a means for normative feedback). The ego-oriented, or strongly extrinsically motivated students, are poor at monitoring formative feedback which has been shown to be an excellent skill in the successful, or "learning oriented", students.

Covington (1983) proposed the "double edged sword" concept to

refer to students who are afraid of real effort because it threatens their ego and, simultaneously, they fear that no effort will get them undesirable attention from the teachers. The crux of the problem is change: planning, introducing, monitoring and re-adjusting to outcomes. Introducing such change requires tact and time. Gradual, mutual, constructive positive feedback (accentuating behaviors to keep) and constructive negative feedback (pointing out behaviors to change), in a quiet and personal atmosphere, provides the interpersonal context and physical setting necessary to execute the many excellent suggestions found in the literature on developmental education and instructional psychology. Perhaps the whole issue may be summarized by active involvement and mutual self-disclosure. If we are to move our schools and classes in the direction of purposeful behaviors, as Pintrich et al. (1993) suggest then, the first step is to admit to ourselves and then to our students this underlying reality:

> "...it is not what the man of science believes that distinguishes him, but how and why he believes it. His beliefs are tentative, not dogmatic; they are based on evidence, not on authority."
>
> Bertrand Russell

Methodology

Objective type tests

The following activities usually take place during a single 50-minute period. Fifteen to eighteen students in class (n=30-35), picked randomly, are given a blank answer sheet and told to randomly circle one answer for each of the 5 matching, 10 multiple-choice questions. These students return completed materials and use the remaining time to study. The other half of the group is given the copy of the test and encouraged to guess if they don't think they know the answers. The questions on the test are randomly selected from each chapter, usually one question for each chapter. Of course, students know that such results will not count towards the final grade. They are reminded and encouraged to provide their "best" guesses in either condition since the discussion will focus on our motivations and outcomes for teaching and learning.

Students then proceed to exchange sheets, if students care to, and to correct the results. The discussion follows immediately the tabulation of scores. The frequency distribution of scores, even with such a small sample of students, number of questions, and a choice from one of four answer stems generates enough

information for quality control. The first item is to show students that the **sensitivity for guessing index.** The average score on the answer sheet for the group of students who made random choices shows this range to be 20 to 30%. We ask students what is to be said of some students who get very low scores (less than 40%) on a test? We suggest that if indeed students with low scores have studied (crammed?), and have the potential to do the course work, they should seek out help with study skills and strategies for learning.

Next, we ask students: "What would it mean if many students got scores in the 20-30% range? We use the answers to introduce the concepts of instrument, instructional, item and interpretation biases. If many students consistently get a question wrong it may be that the wording in the question stem is inadequate. Thus there would be an item (question) bias. If a single wrong answer is consistently chosen then there may be answer type item bias. If many question and answer biases exist on a test then a test (instrument) bias exists. Probably the most common of all instrument biases arises as Canadian college teachers readily adopt textbooks intended for the American markets. Expressions such as college "freshman", "sophomores," "co-eds" etc. do not have their equivalents in Québec and in many parts of Canada.

It may be that the wording of the materials during the course of the teacher's presentation misled students. Ordinarily such instructional biases arise when the teacher's example, usually off-the-cuff, is not in keeping with the theory presented in the lectures or the textbook. In these cases teachers have to recognize that an instructional bias may have occurred[5]. Finally, it may be that individual differences in perception and memory, along with cultural contexts contribute to interpretation biases. For example, students know that cigarette smoking is related to the incidence of lung cancer. If the question stem asks: "What should persons do if they know that there is a -.95 (minus ninety-five) correlation between smoking and lung cancer?", nearly all students will choose the option "stop smoking" rather than the correct one, "increase smoking", because the relationship of the elements in the question stem is negatively correlated[6] or inversely related. If such questions must be included in tests than the teacher must draw student attention to the fact that they are to work with only the information given in the question stem and not with tacit intelligence.

Many students are quick to notice that results are quite similar

---

[5] A teacher is held responsible to see to it that a test matches course contents and objectives.

[6]Negative correlation: "An association between increases in one variable and decreases in another." Wade,C. and Tavris,C. (1993) Psychology 3rd ed. N.Y.: HarperCollins page G-7.

whether students have, or have not, had the questionnaire. The idea is introduced to show that an **instructional index** is possible since having received instruction in course content should translate into higher performances than for students who have not received such instruction. What do students think about the results for students who have had the course, and made a real effort, compared with the results of students who simply were allowed to write the final examination without having taken the course? Asking for a few volunteer students who are not registered for the course to write one of the tests often gives a reasonable estimate of the instructional index. The courageous teacher will ask former students to write the final. I thought I would have to offer a financial incentive to get such help but, to my surprise, staff, teachers and administrators as well as many students volunteered! It has been my experience that the more technical the material the better the instructional index. For example, the instructional index is fair for introduction to psychology and excellent for psychology of mental health. One may more easily discriminate amongst those who have had the mental health course than amongst those who have had the introductory course in psychology.

We find it necessary to introduce the concepts of **discrimination and difficulty indexes** in the context of the discussion about instruction since these topics relate to the current students and their performances. Nothing in testing is more motivating to students than feedback about results. To determine these indexes:

1. Rank test scores in descending (high to low) or ascending order. For example, in descending order, the scores of 20 students on a 33 item multiple choice test with four alternatives is:

94,91,88,85,85,79,76,76,76,73,73,73,73,70,67,64,61,52,39

2. Ideally you should select the top 27% of each group. However, so long as you choose between the range of 25% to 33% you will be okay. With small groups the upper range of 33% is desirable in order to make comparisons between the high and low srores as meaningful as possible.

In the preceding distribution of scores 94 through 85 are the top five high scores and 67 through 39 are the five low scores. Five scores represent 25% of the total sample of 20 students in the example being used here. As the size of each group of the top or bottom students increases, so do the reliability of the indexes. Just remember that with small samples of students (n=10 or less) results are suggestive of changes to be made. If students from different groups take the same test then results could be pooled to increase reliability.

3. Record the responses to each alternative for both the high and low groups. At this point you may wish to prepare a table to facilitate data entry and make eventual comparisons easier. Such a sample table appears on the following page.

4. Calculate the discrimination and difficulty indexes, using these formulae:

DISCRIMINATION INDEX = (High - Low)/N

DIFFICULTY INDEX = (High + Low)/2N

where,

High = the number of students in the high scorers who provided the correct response

Low = the number of students in the low scorers who provided the correct response

N = The size of the sample in the high or low group

In our example in the table on the preceding page we find that for question number 1 all five of the high scorers got the correct answer while only one of the five low scorers did so.

Discrimination Index = (5 - 1) / 5 = 0.80

Generally, discrimination indexes over 0.40 are considered very good. A discrimination index less than 0.30 indicates the question should be revised or dropped. A discrimination index in the 0.31 to 0.39 range indicates minor changes are necessary.

In our example Question 1, the

Difficulty Index = (5 + 1)/10 = 0.60

The difficulty index is interpreted in the light of the midpoint between the sensitivity for guessing index (25% for 4-alternatives, 20% for 5 alternatives and so on) and the highest possible score which is usually 100%. Thus, the maximum discrimination amongst the high and low scorers occurs at the midpoint between this range. In our example for all questions on the test, the ideal difficulty index is (100%-25%)/2 = 62.5%. Our score of 0.60 on item 1 is very close to this number.

High difficulty indexes suggest poor discrimination also. Consider the cases in items 2 and 3 of the table on the next page. Low discrimination scores are associated with either very difficult or

Table 1: Item analysis of a 33 item test with four alternatives for 20 students.
TEST Name: Introduction to Psychology-Group A, Quiz 1

Test Item Number the	Number of correct responses chosen by students in the		Discrimination Index	Difficulty Index	Decision about
	top 25% (N=5)	bottom 25% (N=5)	(High-Low)/N	(High+Low)/2N	test item
#1	5	1	(5-1)/5=0.80	(5+1)/10=0.60	Retain as is
#2	5	4	(5-4)/5=0.20	(5+4)/10=0.90	Too easy
#3	1	0	(1-0)/5=0.20	(1+0)/10=0.10	Too difficult
#4	1	4	(1-4)/5=-0.60	(1+4)/10=0.50	Check for a bias interpretation
.					
.					
.					
#33					

very easy difficulty indexes. An item that does not discriminate well between high and low scores must be replaced since it is also either too difficult or too easy. The teacher may want to consider that the discrimination and difficulty indexes are indicators of problem topics. The idea is not to drop the topic or testing on the topic, but rather to examine if the topic has received enough attention in lectures (an instructional bias); if the wording of the question or answer alternatives (an item bias) is adequate; or, if selective attention and memory are not contributing to students' misperceptions of the question or alternatives (interpretation bias). In a quick review of the item during class you will find that tudents cooperate very well in providing you with such feedback.

Based on these discrimination and difficulty indexes we decide that item 1 on this test did a good job. We now turn our attention to the second and third items on this test. The discrimination and difficulty indexes are poor. In item 2 the discrimination is too low and the difficulty index suggests the item is much too easy. In item 3 the discrimination is equally low but this time the discrimination index reveals the item is much too difficult. Both items 2 and 3 should be re-written. Perhaps there is a problem with the wording, grammatical inconsistencies or give-away; perhaps the distracters are too easy or too close.

The validity of the test (instrument bias) now depends on how many of the items on the rest of the test have discrimination and difficulty indexes similar to item 1 (good item) or to items 2 or 3 (poor items). Often you will find that poor indexes relate to a theme or topic. This is a clear indication of instructional or interpretation biases.

An instrument bias exists when there are fewer questions counting towards the grade that remain than there are questions which are discarded. An item bias arises when the question is really a multiple question or any of other related developmental problems (see Williams,1990). You should drop these items from the test, determine the grade for the test on this revised version, determine the nature of the bias, and then plan on re-testing on this topic. You will find many rewards for such an approach, not the least of which is a dramatic increase in student attention, preparation and performance.

Also, a special case of interpretation bias arises when the group of poorer scorers chooses the correct answer more often than the group of high scorers. Item 4 in table 1,page 17, is an example of the results from a question that ought to be closely examined for an interpretation bias. The discrimination score reveals that poorer students do better than high scorers on this item.

Our multiple-choice tests are timed and vary in length from 18 to

30 questions. The exact number of questions is determined by the length of the question and answer stems (reading and information processing times); the level of difficulty of the questions (factual, applied or conceptual, in order of comprehension difficulty); and, the number and complexity of the major learning objectives for each chapter of assigned reading. Over any one session there are five regularly scheduled, 50-minute quizzes[7]. Following each of the first four quizzes (time doesn't permit a follow-up to the last quiz) one class, in the following week, is used to perform the item analyses. Such procedures can be carried out by the professor ahead of time.

Students are expected to attend class, take lecture notes, participate in class demonstrations, view 2 videos and prepare several labs. The students' grades fall into a rule of thirds: one-third of the grade comes from their video and lab reports, one-third from the quizzes, and one-third from the final examination. The students are told that the final examination is a comprehensive review of the course objectives. They are also told that overlearning that occurs during quiz reviews enhances learning and recall which serve as an excellent preparation for the final exam. The schedule of events and the due dates are stated in the course outline. Students excused from a quiz for a serious reason (medical note or death in the family) are allowed to write a make-up quiz at the end of the term.

Subjective type tests

Much of what has been said in the preceding discussion could apply to a discussion about subjective type tests. Essays, brief position papers, term papers, library research, reports of all types etc. immediately come to mind. A teacher could examine in-class student essays, for example, to uncover their understanding of basic terminology, application of information processing rules, or generating original thinking. In the process the teacher could count the number of students who fail an essay question and compare results with the reading difficulty level of the question. Perhaps the vocabulary is suitable only for the better student. So, in this respect, content analysis of essay type work has some roots in techniques suited for objective type evaluations.

Whatever the type, subjective tests have three items in common: the student must demonstrate some mastery of the content as well as

---

[7] However it may be possible to schedule only a mid-term and final examination. Students could have opportunities for correcting mistakes if the weight of the mid-term is slight and the weight for the final is heavy. We assume then that the comprehensive mid-term exam includes detailed formative feedback to students about metacognitions for studying and to teachers about instrument, item, instructional and interpretation biases.

intellectual abilities and skills; the subjectivity of the evaluation process must be minimized, or at least rationalized; and, the preference for the "power" of a test (over "speed[8]") is emphasized.

A widely used system is based on Bloom et al's classic Taxonomy of Educational Objectives (1956). The taxonomy is a "content by objectives" approach to understand how well the student demonstrates the acquisition of knowledge and the use of intellectual abilities and skills to convey that knowledge. A series of hierarchical objectives are presented in the cognitive domain to measure student understanding and use of: knowledge, comprehension, analysis, synthesis and evaluation. The same hierarchical structure exists for the affective domain (Krathwohl et al, 1964) which measures how well students work with values, attitudes, interests, opinions, beliefs and appreciation. Student work is measured by their ability to receive (attend) respond, value, organize and determine structure from what they read and write.

Such an approach became popular because it provided teachers with help in rationalizing their feedback. Of course, such an approach was not consistent with maladaptive learning behaviors. The Taxonomic approach assumes students are active participants who want and use formative feedback to improve their performances. The psychology of individual differences, the essense of psychological testing, and the emergence of school and educational psychology soon established that there was more to teaching and learning than formative feedback/evaluation. Normative, summative and diagnostic feedback/evaluation processes soon became a part of teachers' daily realities. The diagnostic feedback brought with it the realization that reporting on performances could be based on domain, criterion or normative references.

It is relatively easy then to understand how these complexities have created and supported the variety of schemes for measuring and evaluating students' written work. Whatever position the teacher favors, we propose that the student who is informed of the context and process of evaluation and is provided inputs into that process will become more motivated since s/he is actively involved in the evaluation of the product of teaching/learning. This approach is by no means novel. We have known for decades that student involvement produces superior learning results and higher classroom motivation. What teachers have been reluctant to do is to involve students in testing procedures. We are not advocating turning control over to students. Allow them inputs and share with them the reasons and

---

[8] "Power" and "speed" in testing refer, respectively, to the fact that the person writing the test may have virtually unlimited time to formulate his/her answers, or be faced with time constraints. In our North American society, we most often find a mixture of power/speed for classroom testing and tend more towards power for essay type work.

motives that guide your decision for one type of feedback and/or evaluation.

There are excellent references in educational testing, educational psychology and measurement and evaluation textbooks. Perhaps one reference that helps the teacher think through the process measurement and evaluation is Martuza's (1977) Aplying Norm-Rerferenced and Criterion-Referenced Measurement in Education. To understand the internalized rules and norms teachers develop is well presented in Airasian and Madaus's "Criterion-Referenced Testing in the Classroom" (Martuza, 1977; pages 330-344).
The Center for the Study of Evaluation at the University of California at Los Angeles ("UCLA") and the National Center for Research to Improve Postsecondary Teaching and Learning ("NCRIPTAL") at the University of Michigan (Ann Arbor) have publications and tools to help teachers help students with respect to measurement and evaluation. A very readable and practical reference for teachers is Cross and Angelo's (1988) Classroom Assessment Techniques - A Handbook for Faculty.

Results

Both the academically weak and strong students attend these 1-2 hour workshops held during the week following the quiz. In all, there are four such workshops and the trend is for an increasing number of students to attend from one workshop to the next during a session. Anywhere from one-fifth to nearly one-half of all students attend the test review workshop to understand where they made the mistake and how to avoid it on future tests. Additionally, from 10% to 20% of all students, or about 40% of all students who attend the workshops, make appointments for personal reviews. Students rarely come to bicker about grades. They come by to discuss better learning strategies and work habits. Our College has adopted Fraser's (1993) Making Your Mark, a 36 page guide to facilitate students' coping with academic life and the teacher's academic advising role.

Course/teacher evaluations show students are very satisfied with the teacher's attitudes and behavior with respect to testing and evaluations. The number of requests for re-reads was low but now they have disappeared. Students understand how they made their mistakes. They realize the teacher is willing to help them learn how to avoid making mistakes by developing better learning strategies and study habits---as opposed to coaching them to pass tests! The grade becomes less of an issue. It appears that there are indications of decreasing performance and evaluation anxieties.

Discussion

Students learn the contingency between real effort and results. Each student participates to create his or her own reinforcement history. Students learn to think for themselves, to set realistic goals, the means for attaining them, and to accept the ultimate responsibility for attaining their academic potential. This is facilitated to the degree that students not only hear the teachers tell them this but see the teachers engage in it. This spirit has been wisely expressed in the following Ancient Chinese Proverb which has been adapted in a poster for the Classmate DIALOG services:

Tell me,
I forget.
Show me,
I remember.
Involve me,
I understand.

These are results based on clusters of convenience. Generalizations hopefully, but not scientifically, are that the processes one has engaged in will be valid from teacher to teacher. How much gain there is to be expected ought to show much variability in the types and quality of changes. However, it does appear that constructive changes "assessment for improvement" is possible. After all, students don't expect us to be perfect but perhaps to show some concern about helping them know when and how to deal with the relatively large amounts of knowledge teachers ask them to process.

References

Anderson,S.B. (1987) The Role of the Teacher-Made Test in Higher Education. In D.Bray and J.Belcher (Editors) Issues in Student Assessment. **New Directions for Community Colleges**, (no.59,pages 39-44) San Francisco: Jossey-Bass.

Argyris,C. and Schön,D.A. (1974) Theory in Practice: Increasing Professional Effectiveness San Francisco: Jossey-Bass.

Association québécoise de pédagogie collégiale (1994,juin) Évaluation! Évolution? Où s'en va le collégial? ("The Evolution of Evaluation: Where are our colleges heading?") 14e colloque, Québec.

Bateman,D. (1989) A Longitudinal Study of the Cognitive and Affective Growth of Cégep Students Champlain Regional College, St-Lambert-Longueuil Campus, St-Lambert, Québec, J4P 3P2.

Bloom,B.S. (1956) (Ed.) Taxonomy of Educational Objectives, Handbook I: Cognitive Domain N.Y.: McKay.

Covington,M.V. (1983) Motivated Cognitions. In S.G. Paris, G.M.Olson, and H.W. Stevenson (Editors) Learning and Motivation in the Classroom. Hillsdale:New Jersey. Lawrence Erlbaum Associates.

Cross, K.Patricia (1988) Feedback in the Classroom: Making Assessment Matter. **The American Association for Higher Education** (AAHE), One Dupont Circle, Suite 600, Washington,D.C. 20036.

Cross, K.Patricia and Angelo,T.A. (1988) Classroom Assessment Techniques - A Handbook for Faculty National Center for Research to Improve Postsecondary Teaching and Learning ("NCRIPTAL") The University of Michigan, Ann Arbor.

Ellsworth,R.A., Dunnell,P. and Duell,O.K. (1990) Multiple-Choice Test Items: What Are Textbook Authors Telling Teachers? **The Journal of Educational Research**, May/June 83(5) 289-293.

Fraser,L. (1993) Making Your Mark, 3rd ed. LDF Publishing, 16151 Old Simcoe Road; Port Perry, Ontario L9L 1P2, 36 pages.

Kirby,P.C. and Teddlie,C. (1989) "Development of the Reflective Teaching Instrument", **Journal of Research and Development in Education**, 22(4), Summer, 45-51.

Krathwohl,D.R., Bloom,B.S. and Masia,B.B. (1964) Taxonomy of Educational Objectives II: Affective Domain N.Y.: McKay.

MacCuish,D.A. (1986) "The course development model in higher education: Improving tests and instruction." ERIC ED273-169.

Marso,R.W. and Pigge,F.L. (1992, April 20-24) A Summary of Published Research: Classroom Teachers' Knowledge and Skills Related to the Development and Use of Teacher-Made Tests. Paper presented at the Annual Meeting of the American Educational Research Association,San Francisco.

Martuza,V.R. (1977) Applying Norm-Referenced and Criterion-Referenced Measurement in Education Boston: Allyn & Bacon.

Milton,O. (1982) Will That Be on the Final? Springfield,Illinois: Thomas.

Pintrich,P.R.,Cross,D.R., Kozma,R.B. and McKeachie,W.J. (1986) "Instructional Psychology." In **Annual Review of Psychology** 37,611-651.

Pintrich,P.R. and Johnson,G.R. (1990) Assessing and Improving Students' Learning Strategies. In M. D. Svinicki (Ed.) The Changing Face of College Teaching, **New Directions for Teaching and Learning**, (No. 42, Summer, pages 83-92) San Francisco: Jossey-Bass.

Pintrich,P.R., Marx,R.W. and Boyle,R.A. Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. **Review of Educational Research** 63(2), 167-199.

Pollio,H.R. and Humphreys,W.L. (1988) Grading Students. In J.H. McMillan (Ed.) Assessing Students' Learning **New Directions for Teaching and Learning**, (No.34,pages 85-97) San Francisco: Jossey-Bass.

Rubadeau,D., Garrett,Wm.A. and Rubadeau,R.J. (1990) Appropriate Testing College of New Caledonia Press, 3330-22nd Ave, Prince George, British Columbia V2N 1P8.

Schön,D.A. (1983) The Reflective Practioner San Francisco: Jossey-Bass.

Stiggings,R.J. (1988) Revitalizing Classroom Assessment: The Highest Instructional Priority. **Phi Delta Kappan**, January, 363-368.

Stiggins,R.J. and Brdigeford,N.J. (1985) The Ecology of Classroom Assessment. **Journal of Educational Measurement** 22(4) 271-286.

Stiggins,R.J., Conklin,N.F. and Bridgeford,N.J. (1986) Classroom Assessment: A Key to Effective Education. **Educational Measurement: Issues and Practices**, Summer, 5-17.

Talbot,G.L. (1994) The Assessment of Student Study Skills and Learning Strategies to Prepare Teachers for Academic Advising Tasks. Available from the author (10$) c/o: Champlain-Saint Lawrence, 790 Nérée Tremblay Street, Ste-Foy, Province of Québec, G1V 4K2, 85 pages.

Taylor,H, Greer,R.N. and Mussio,J. (1978) Construction and Use of Classroom Tests: A Resource Book for Teachers Learning Assessment Branch, Ministry of Education, Province of British Columbia, December, ERIC ED190-608.

Thomas,M.D. (1988) Test Item Analysis: Analyzing your test results. In D.M. Stein Instructor's Manual to Accompany R.R. Bootzin and J.R. Acocella Abnormal Psychology: Current Perspectives (5th ed.) N.Y.:Random House, pages "a" thru "g".

Williams,Jane M. (1991) Writing Quality Teacher-Made Tests: A Handbook for Teachers. Available from the author (5$ US) c/o: Wheaton High School, 12601 Dalewood Dr., Wheaton, Maryland 20906-4168. ERIC ED349-726.